Cross-cultural comparability of noncognitive constructs in TIMSS and PISA

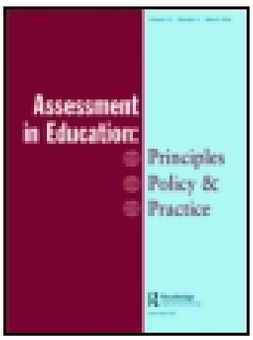
He, Jia; Barrera-pedemonte, Fabián; Buchholz, Janine

Published in:
Assessment in Education: Principles, Policy & Practice

DOI:
10.1080/0969594X.2018.1469467

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record



Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: https://www.tandfonline.com/loi/caie20

Cross-cultural comparability of noncognitive constructs in TIMSS and PISA

Jia He, Fabián Barrera-Pedemonte & Janine Buchholz

To cite this article: Jia He, Fabián Barrera-Pedemonte & Janine Buchholz (2018): Cross-cultural comparability of noncognitive constructs in TIMSS and PISA, Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2018.1469467

To link to this article: https://doi.org/10.1080/0969594X.2018.1469467

Published online: 24 Jul 2018.

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE
https://doi.org/10.1080/0969594X.2018.1469467

Routledge
Taylor & Francis Group

Cross-cultural comparability of noncognitive constructs in TIMSS and PISA

Jia He[a,b], Fabián Barrera-Pedemonte[c] and Janine Buchholz[a]

[a]German Institute for International Educational Research, Frankfurt am Main, Germany; [b]Department of Methodology and Statistics, Tilburg University, Tilburg, the Netherlands; [c]Center for Advanced Research in Education, Universidad de Chile, Santiago, Chile

ABSTRACT

Noncognitive assessments in Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study share certain similarities and provide complementary information, yet their comparability is seldom checked and convergence not sought. We made use of student self-report data of Instrumental Motivation, Enjoyment of Science and Sense of Belonging to School targeted in both surveys in 29 overlapping countries to (1) demonstrate levels of measurement comparability, (2) check convergence of different scaling methods within survey and (3) check convergence of these constructs with student achievement across surveys. We found that the three scales in either survey (except Sense of Belonging to School in PISA) reached at least metric invariance. The scale scores from the multigroup confirmatory factor analysis and the item response theory analysis were highly correlated, pointing to robustness of scaling methods. The correlations between each construct and achievement was generally positive within each culture in each survey, and the correlational pattern was similar across surveys (except for Sense of Belonging), indicating certain convergence in the cross-survey validation. We stress the importance of checking measurement invariance before making comparative inferences, and we discuss implications on the quality and relevance of these constructs in understating learning.

ARTICLE HISTORY
Received 24 October 2017
Accepted 6 April 2018

KEYWORDS
PISA; TIMSS; measurement invariance; cross-survey validation

Introduction

The Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) of Grade 8 are two flagship large-scale educational surveys. They provide international comparative data of students approaching to or transitioning up from the end of lower secondary education for research and evidence-based policy-making. Despite differences in target populations (15-year olds vs. eighth graders), assessment frameworks (skill-based vs. curriculum-based), test characteristics (e.g. lengths

CONTACT Jia He j.he2@tilburguniversity.edu

Present affiliation for Fabián Barrera-Pedemonte is Universidad San Sebastian, Facultad de Ciencias de la Educacion (address: Bellavista 7, Recoleta, Santiago de Chile, postal code: 8420524).

of test and foci of context questionnaires) and scaling methodologies (i.e. different item response theory models), they share a number of similarities and provide complementary information (Michael, Ina, Alka, & Corinna, 2014; OECD, 2015). So far, comparisons of the two surveys have targeted cognitive assessment (e.g. Klieme, 2016; Neidorf, Binkley, Gattis, & Nohara, 2006; Wu, 2010). Much less attention has been given to the complementary potential of noncognitive assessments (i.e. the context questionnaires) in the two surveys, which is a missed opportunity to borrow strengths from each survey, enhance methodological rigour for such noncognitive assessments and provide validity of the statistical basis of policy arguments.

The foci and terminologies may differ in constructs measured in the noncognitive assessments of the two surveys, yet both aim to provide insight into contexts and factors related to learning. These assessments show a fair degree of similarities in the measured constructs. In the 2015 assessments of PISA and TIMSS, many constructs in the context questionnaires administered to students, teachers and principals show an overlap in the theoretical concepts and item wording. For instance, in the student assessments, Likert scale items on instrumental motivation, enjoyment of science and sense of belonging to school were administered and scale scores were produced in both surveys. Apparently, these are factors recognised by both surveys as important in understanding student learning (Mullis & Martin, 2013; OECD, 2015).

Instrumental motivation and enjoyment of science are two facets of students' motivation associated with science achievement (House, 2004; OECD, 2016; Wigfield, Eccles, & Rodriguez, 1998). The former reflects the desire of students to learn science as a mean to obtain related rewards in the future, such as improving career opportunities and selecting into scientific fields of study (Nagengast & Marsh, 2014). Enjoyment of science refers to the intrinsic drive for science learning and the satisfaction with this subject itself (Ryan & Deci, 2009). The sense of belonging to school is a dimension of students' well-being comprising feelings of social acceptance and attachment to the school community (Baumeister & Leary, 1995).

Research conducted in the last decade has indicated that instrumental motivation (Yu, 2012), enjoyment of science (Cosgrove & Cunningham, 2011; Grabau & Ma, 2017; Lam & Lau, 2014) and sense of belonging to school (Chiu, Chow, McBride, & Mol, 2016; Topçu, Erbilgin, & Arikan, 2016), as measured in TIMSS or PISA, are positively associated with science achievement in developed countries. Nonetheless, important variation across cultures is repeatedly reported regarding the size of this relationship. For instance, Yu (2012) found that instrumental motivation predicted TIMSS science scores of eighth-grade students in the US, but no evidence of such a relationship was found among their East Asian counterparts. Ainley and Ainley (2011) also reported that whilst enjoyment of science contributed to science achievement across the 57 cultures of PISA 2006, such a relationship varied depending on the cultural traditions of school systems.

This paper integrates context questionnaire and achievement data from the 2015 PISA and 2015 TIMSS to illustrate a case of cross-validating the assessments of these three noncognitive constructs – i.e. instrumental motivation, enjoyment of science and sense of belonging to school – within and across surveys. In the following, we first present a methodological framework of bias and equivalence which guides the test of data comparability, and then we review how bias and equivalence are taken into consideration in the noncognitive assessment in each survey.

Methodological rigour in cross-cultural comparisons: bias and equivalence

For both cognitive and noncognitive assessments, the presence of bias indicates that scores from the assessment in different cultures[1] reflect some cultural characteristics other than what the assessment is intended to measure. Before any comparative inference is made, bias needs to be detected and ruled out. According to van de Vijver and Leung (1997), three levels of measurement bias can be distinguished: (1) Construct bias: the construct that is the target of the assessment has a different meaning in different cultures; (2) Method bias: there is incomparability due to differences in sampling, respondents' use of the test instruments and administration modes; and (3) Item bias: an item has a different meaning in different cultures.

Conceptual and psychometric comparability should be demonstrated before any comparative inferences are made (van de Vijver & Leung, 1997, 2000). After data are collected, a host of psychometric tools can be used to check the comparability (absence of bias) (Boer, Hanke, & He, in press). Levels of comparability (also called invariance) across groups can be demonstrated either in the framework of confirmatory factor analysis or in the item response theory framework (IRT: using logistic models) (Reise, Widaman, & Pugh, 1993). Multigroup confirmatory factor analysis (MGCFA) is currently the 'most widely used' invariance testing method (Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014, p. 1). Following the terminology used in MGCFA models, we can distinguish three levels of invariance (van de Vijver & Leung, 1997): (1) *Configural invariance* indicates that items measuring a construct cover facets of this construct adequately. In statistical terms, this means that items in a measure exhibit the same configuration of salient and non-salient factor loadings across groups; (2) *Metric invariance* means that the items measuring a construct have the same factor loadings across groups (in IRT models, this means the item discriminations are equal). With metric invariance satisfied, associations between variables can be compared across groups. For instance, the correlation between student motivation and achievement in each culture can be computed and compared across cultures, if both student motivation and achievement prove to be metric invariant. (3) *Scalar invariance* implies that items have the same loadings (i.e. item discrimination in IRT) and intercepts (i.e. item difficulty parameters in IRT) across groups. This would mean that score levels on scalar-invariant measures reflect the same psychological aspect across cultures; for example, two individuals from two different cultures who both have a mean score of 4 on a motivation scale can be regarded as equally motivated, whereas a student exhibiting a value of 2 can be regarded less motivated than a student with a scale score of 2. Only with scalar invariance can such scale scores be validly compared across cultures.

Bias assessment in PISA and TIMSS

As mentioned before, the conceptual demonstration of invariance for both the cognitive and noncognitive assessments occurs before data are collected, through extensive study of the literature to operationalise a concept (to address construct bias), sampling of adequate items to capture the construct and carefully translating and adapting item content (to address method and item bias) to ensure comparability and ecological validities. In both PISA and TIMSS, such steps are taken during the preparation of the assessment frameworks, extensive

field trials and national adaptations (Klieme & Kuger, 2015; Kuger, Klieme, Jude, & Kaplan, 2016; Mullis & Martin, 2013).

During data analysis, both surveys have made great effort in minimising bias in the cognitive assessment through identifying items with differential item functioning and making adjustments in the estimation of cognitive ability scores (Mullis & Martin, 2013; OECD, 2017). Yet, the measurement bias issues in the noncognitive assessment are less extensively treated compared with the cognitive assessments in these surveys (Braeken & Blömeke, 2016). Compared with the cognitive assessments, noncognitive assessments using self-report Likert scales seem more vulnerable in providing comparable scores. With a large number of cultures involved (such as in large-scale international assessments) and so many culture-specific variations (such as conceptual differences in constructs), respondents' interpretations of item content and cultural and personal preferences in using response scales, it is difficult to obtain high levels of invariance across all cultures (He & Kubacka, 2015). Whilst the international technical report of either PISA or TIMSS provided information on the formal testing of strict measurement invariance, the scale scores used in the international comparisons were based on different assumptions and scaling methods (as detailed below). Despite the increased awareness to examine data comparability in recent years (Vandenberg & Lance, 2000), applied researchers tended to make use of the scale scores reported in the international databases of PISA and TIMSS for substantive topics (e.g. educational evaluations), where measurement invariance testing for scales of interest in targeting cultures was seldom included as an integral part of the analysis (e.g. Chiu & Zeng, 2008). There is a need for applied researchers to have a thorough understanding on the scaling methodologies and possible drawbacks of these reported scale scores.

According to the international technical reports, both surveys reported internal consistency (using Cronbach's Alpha) and factor loadings (with Principal Component Analysis) to validate noncognitive scales. For TIMSS 2015, a one-parameter IRT modelling approach (i.e. the partial credit model) was used to calibrate the combined data from all cultures and estimate scale scores. Here, there is an implicit assumption of invariance of item parameters across cultures in TIMSS (Martin et al., 2016a). In contrast, PISA 2015 used a two-parameter IRT modelling approach (i.e. the generalised partial credit model) to examine the comparability of scale items and, instead of constraining all item parameters to be equal, allowed for country/language group-specific item parameters when they exhibited bad item fit (OECD, 2017). The recognition of cultural variations in measurement and the flexibility in treating such variations in PISA are a step forward in better estimating scale scores of noncognitive constructs.

The present study

As the two surveys used different items to measure the noncognitive constructs, applied different scaling methods to derive scale scores and treated statistical bias assessment in different manners, a straightforward comparison of similar constructs in noncognitive assessments across surveys is hard to achieve. Indeed, a review of literature reveals scarcely simultaneous use of noncognitive data from PISA and TIMSS. With overlapping constructs targeted in both surveys in 2015, a cross-validation is possible, and the most comparable scales can be sought.

Therefore, the present study aims to (1) demonstrate the measurement comparability of noncognitive student-level constructs targeted in both surveys across cultures in a common framework (i.e. MGCFA), (2) check convergence of different scaling methodologies (i.e. IRT and MGCFA) on deriving scale scores within each survey and (3) check the convergence between surveys (with a common scaling method). The results are expected to provide solid evidence on the level and extent of comparability of noncognitive scales in both surveys, before these data are used for various comparative research projects and for policy feed.

Method

We make use of the 2015 PISA and 2015 TIMSS (eighth grade) survey to check convergence of student self-reported constructs from the respective student context questionnaires. Only the 29 cultures that participated in both surveys are included for comparisons.

Measures

Targeted constructs with similar item wording include students' instrumental motivation (data of 29 cultures available in both surveys), enjoyment of science (21 cultures) and sense of belonging to school (28 cultures). These constructs are operationalised in each survey by asking students to rate a set of items on a four-point Likert-type scale, which is designed to collect their level of agreement with each statement (PISA: 'strongly disagree', 'disagree', 'agree', 'strongly agree'; TIMSS: 'disagree a lot', 'disagree a little', 'agree a little' and 'agree a lot'). Table 1 shows the items administered in each survey respectively.

Despite the acknowledged differences between assessments, including that TIMSS in general uses more items, that there is no match in the name given to scales and that some items address supplementary aspects of the construct, there is conceptual overlap based on how these concepts are defined in the assessment frameworks and similarity in the wording of several items (Mullis & Martin, 2013; OECD, 2015). Take the construct enjoyment of science for example: all the 5 items from PISA address the majority of aspects included in the TIMSS scale. These elements cover experiences of having – or not having – fun with science (PISA: 'I generally have fun when I am learning < broad science > topics'; TIMSS: 'Science is boring'), assertions on the liking for this subject (PISA: "I like reading about < broad science > "; TIMSS: 'I like science'), the enjoyment with learning activities (PISA: 'I am happy working on < broad science > topics'; TIMSS: 'I like to conduct science experiments'), with studying – or not studying – science in general (PISA: 'I enjoy acquiring new knowledge in < broad science >'; TIMSS: 'I wish I did not have to study science'), as well as the growing interest in the subject (PISA: 'I am interested in learning about < broad science >'; TIMSS: 'I learn many interesting things in science').

The reliability of the six scales was generally high within participating cultures. For example, across cultures that took part in TIMSS, Cronbach's alpha coefficients showed values above .86 for instrumental motivation, .81 for enjoyment of science and .64 for sense of belonging to school (only 9 out of the 46 cultures yielded values below .80) (Martin et al., 2016b). In the case of PISA, these values were above .88 for instrumental motivation, above .79 for enjoyment of science and .71 for sense of belonging to school (only Belgium, France and Korea showed reliabilities below .80) (OECD, 2017). These results suggested that the scales selected in this paper have high levels of internal consistency within cultures.

Table 1. Items and scales used in PISA 2015 and TIMSS 2015 for the constructs: instrumental motivation, enjoyment of science and sense of belonging to school.

Construct	PISA 2015	TIMSS 2015
Instrumental motivation	Scale: Instrumental motivation 1. Making an effort in my <school science> subject(s) is worth it because this will help me in the work I want to do later on 2. What I learn in my <school science> subject(s) is important for me because I need this for what I want to do later on 3. Studying my <school science> subject(s) is worthwhile for me because what I learn will improve my career prospects 4. Many things I learn in my <school science> subject(s) will help me to get a job	Scale: Students value science 1. I think learning science will help me in my daily life 2. I need science to learn other school subjects 3. I need to do well in science to get into the <university> of my choice 4. I need to do well in science to get the job I want 5. I would like a job that involves using science 6. It is important to learn about science to get ahead in the world 7. Learning science will give me more job opportunities when I am an adult 8. My parents think that it is important that I do well in science 9. It is important to do well in science
Enjoyment of science	Scale: Enjoyment of science 1. I generally have fun when I am learning <broad science> topics 2. I like reading about <broad science> 3. I am happy working on <broad science> topics 4. I enjoy acquiring new knowledge in <broad science> 5. I am interested in learning about <broad science>	Scale: Students like learning science 1. I enjoy learning science 2. I wish I did not have to study science 3. Science is boring 4. I learn many interesting things in science 5. I like science 6. I look forward to learning science in school 7. Science teaches me how things in the world work 8. I like to conduct science experiments 9. Science is one of my favourite subjects
Sense of belonging to school	Scale: Sense of belonging to school 1. I feel like an outsider (or left out of things) at school 2. I make friends easily at school 3. I feel like I belong at school 4. I feel awkward and out of place in my school. 5. Other students seem to like me 6. I feel lonely at school	Scale: Students' sense of school belonging 1. I like being in school 2. I feel safe when I am at school 3. I feel like I belong at this school 4. I like to see my classmates at school 5. Teachers at my school are fair to me 6. I am proud to go to this school 7. I learn a lot in school

Notes: Information compiled from OECD/PISA student questionnaire (2017) and IEA/TIMSS 8th grade science student questionnaire (2017).

Analysis strategies

Three sets of analysis were performed. The first involved measurement comparability checks and the second and third analyses involved comparisons of correlations between the six non-cognitive scales and science achievement data. We first statistically checked measurement invariance at different levels (configural, metric and scalar) in the framework of MGCFA (Vandenberg & Lance, 2000; Wu, Li, & Zumbo, 2007) for each construct in each survey (for each pair of constructs, only cultures with data in both surveys were included to ensure that the model fit comparisons are based on the same set of cultures). We treated data as continuous and used the full information maximum likelihood estimation. We also made use of the senate weights calculated in each survey, which rescaled sample sizes to be fixed at 500 cases per country for both surveys. The use of this weighting factor is recommended to balance the contribution of each country in the estimation to avoid the final solution being pulled towards countries with larger sample sizes (Stapleton, 2014). This analysis was executed with Mplus 7.3 (Muthén & Muthén, 1998–2011).

Second, for scales with measurement comparability established (e.g. metric invariance), the factor scores were estimated in the MGCFA models and they were compared with the scale scores reported in the international database (based on respective IRT models) in each survey separately. This serves to check the robustness of scaling methods within a survey. Next, the correlations of the scale scores with science achievement per culture (based on the factor scores estimated from the MGCFA models) were compared across surveys.

Results

We report the results in three parts: the measurement invariance testing of the six scales (three in each survey), the comparisons of factor scores within surveys and the comparisons across surveys.

Measurement invariance

The measurement invariance of each scale in each survey was checked in MGCFA models. Model fit in MGCFA was evaluated by three measures, including χ^2 statistics (although it is rather sensitive to sample sizes), Comparative Fit Index (CFI: above .90 considered acceptable) and Root Mean Square Error of Approximation (RMSEA: below .08 considered acceptable). The acceptance of a more restrictive model was based on the change of CFI and RMSEA. In the contexts of large-scale assessments, Rutkowski and Svetina (2014) proposed to set the cut point of ΔCFI to .02 and that of ΔRMSEA to .03 from configural to metric models, and both ΔCFI and ΔRMSEA to .01 from metric to scalar models when data are treated as continuous using the maximum likelihood estimator. We followed these guidelines. Table 2 presents the model fit for each scale in each survey.

Based on the changes of CFI and RMSEA, all three scales in TIMSS reached metric invariance (ΔCFI from .01 to .02, and ΔRMSEA = .00 from configural to metric model). Instrumental motivation in PISA showed metric invariance and enjoyment of science only showed configural invariance based on the change of CFI (ΔCFI = .05), and acceptable metric invariance based on change of RMSEA (ΔRMSEA = .00). The configural model of sense of belonging in PISA failed to converge,[2] and the metric model showed rather poor model fit. None of the scales reached scalar invariance. Despite the different numbers of

Table 2. Tests of measurement invariance for the constructs instrumental motivation, enjoyment of science and sense of belonging to school in PISA 2015 and TIMSS 2015.

		PISA				TIMSS			
Construct		χ^2	df	RMSEA	CFI	χ^2	df	RMSEA	CFI
Instrumental motivation (29 cultures)	Config	160.46**	58	.02	.95	35,814.08**	783	.09	.91
	Metric	286.88**	142	.01	.93	41,778.27**	1007	.09	.89
	Scalar	463.50**	226	.01	.89	59,816.27**	1231	.10	.85
Enjoyment of science (21 cultures)	Config	350.71**	105	.02	.93	14,443.19**	546	.06	.96
	Metric	581.89**	185	.02	.88	18,765.94**	706	.06	.95
	Scalar	809.48**	265	.02	.83	3278.85**	866	.08	.92
Sense of belonging to school (28 cultures)	Config	Non-convergence				7961.85**	392	.06	.96
	Metric	1011.39**	303	.02	.70	11,073.28**	554	.06	.94
	Scalar	1190.54**	438	.02	.68	30,511.18**	716	.09	.83

Notes: Error terms of negatively worded items were intercorrelated to account for the wording effects in PISA sense of belonging to school, and in enjoyment of science in TIMSS. RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index.

items in these three pairs of constructs (PISA in general has fewer numbers of items than TIMSS), the model fit indices suggested that, except for sense of belonging to school and enjoyment of science in PISA, the constructs were understood as the same across cultures, and the items were additionally equal reflections of the construct. This indicated that the correlations of these scales could be compared, whereas the scale mean scores were not comparable across cultures.

Correlations with science achievement per culture within each survey

With metric invariance of the scales established, correlational analysis among metric invariant constructs and achievement can be considered valid. Despite the marginal support of metric invariance for enjoyment of science and lack of such support for sense of belonging to school in PISA, we estimated factor scores of all the six scales in the metric invariance model. The measurement invariance of the cognitive assessments of the two surveys is not being tested in the current study, but the technical reports of either survey have demonstrated comparability of achievement scores across cultures, so we used them to explore the convergence of the target scales.

For each target construct in each survey, the correlations of the MGCFA factor scores with science achievement in each culture were compared with the correlations of the factor scores reported in the international database (based on respective IRT models). Figures 1–3 illustrate the associations for the three constructs, respectively. As both surveys reported multiple plausible values for achievement, correlations were calculated for each plausible value and their average was used to create the scatterplots (Figures 1–3). The same procedure was applied for Figures 4–6. In general, the correlations between each scale and achievement scores were positive at the individual level, indicating that instrumental motivation, enjoyment of science and sense of belonging to school were associated with better achievement in science in each culture. All correlations of the correlations were as high as .99 across the scaling methodologies, except for sense of belonging to school in PISA which correlated at .73 (note the poor measurement property of this scale). In general, this pointed to the robustness of findings based on the MGCFA and the IRT scaling.

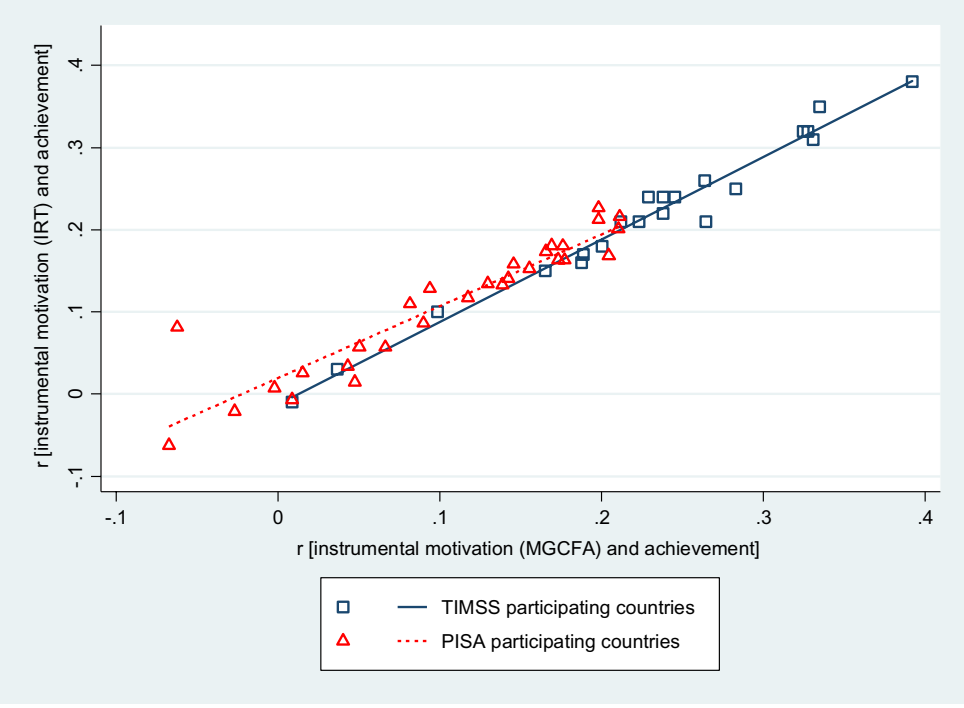


Figure 1. Relationship between within-culture correlations of science achievement and instrumental motivation, where instrumental motivation was scaled based on the MGCFA metric invariance model (*x*-axis) and IRT-based survey reported scale scores (*y*-axis), respectively.

Correlations with science achievement per culture across surveys

The correlations of the science achievement and the scale scores of the three constructs in each culture (as scaled in the MGCFA metric invariance model) are displayed in Figures 4–6, respectively. The *x*-axes display the correlation coefficient between science achievement and the scale score used in TIMSS; the *y*-axes show the corresponding coefficient in PISA; and dots represent cultures with data available in both assessments.

The overall strength of the association was moderately small, yet similar across surveys. All three pairs of scales were positively associated with students' scores in science regardless of the survey to which students responded, yet the size of these associations differ across cultures.

Figures 4 and 5 depicted clear patterns of convergence between surveys for the instrumental motivation and enjoyment of science scales. For both scales, within-culture correlations with science achievement were positively and highly correlated with the corresponding coefficients yielded in the counterpart survey. In other words, cultures with greater associations between the noncognitive scale and science achievement in TIMSS also tended to show the same pattern in PISA. The positive trend in both scatterplots reflected a strong correlation at the culture level (Pearson's $r = .70$ and .76, respectively). Along this line, there is also a recurrent pattern of countries showing lower (Chile and Argentina) and higher (Korea, Australia and Ireland) correlations with achievement, both in instrumental motivation and enjoyment of science.

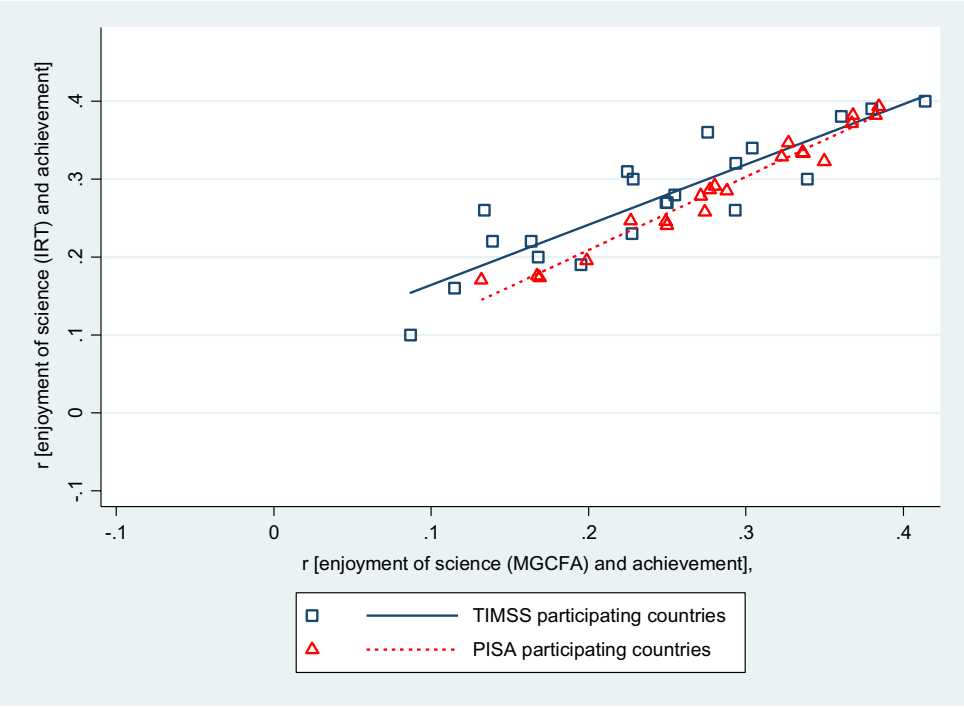


Figure 2. Relationship between within-culture correlations of science achievement and enjoyment of science, where enjoyment of science was scaled based on the MGCFA metric invariance model (*x*-axis) and IRT-based survey reported scale scores (*y*-axis), respectively.

In contrast, results hardly provided evidence of convergence for the sense of belonging to school scale. In this instance, the size of within-culture associations with science achievement tended to increase in one survey and slightly decrease in the other, yielding a negative country-level correlation coefficient (Pearson's $r = -.35$). Although this result may indicate divergence between surveys, the analysis must be put in the context of the generally narrow range of the associations found in PISA compared with TIMSS. This is an interesting finding because it casts doubt on the sensitivity of the PISA scale to capture enough variation with students' achievement in science. At this point, it is worth recalling that this scale did not converge at the configural level of measurement invariance, which implies poor evidence of the capacity of these items to elicit the same construct across cultures.

Discussion and conclusion

Noncognitive outcomes, such as students' motivation and their sense of belonging in school, are equally important as cognitive outcomes because these personal qualities and "soft" skills help students to be successful in school and eventually in society at large (Kuger et al., 2016). This speaks to the importance of reliably and validly measured noncognitive constructs in these surveys. We set out to link the similarities of noncognitive assessments in PISA and TIMSS and cross-validate the similar student self-report scales in either surveys. Using data of three student scales from 29 overlapping cultures, we found that instrumental motivation

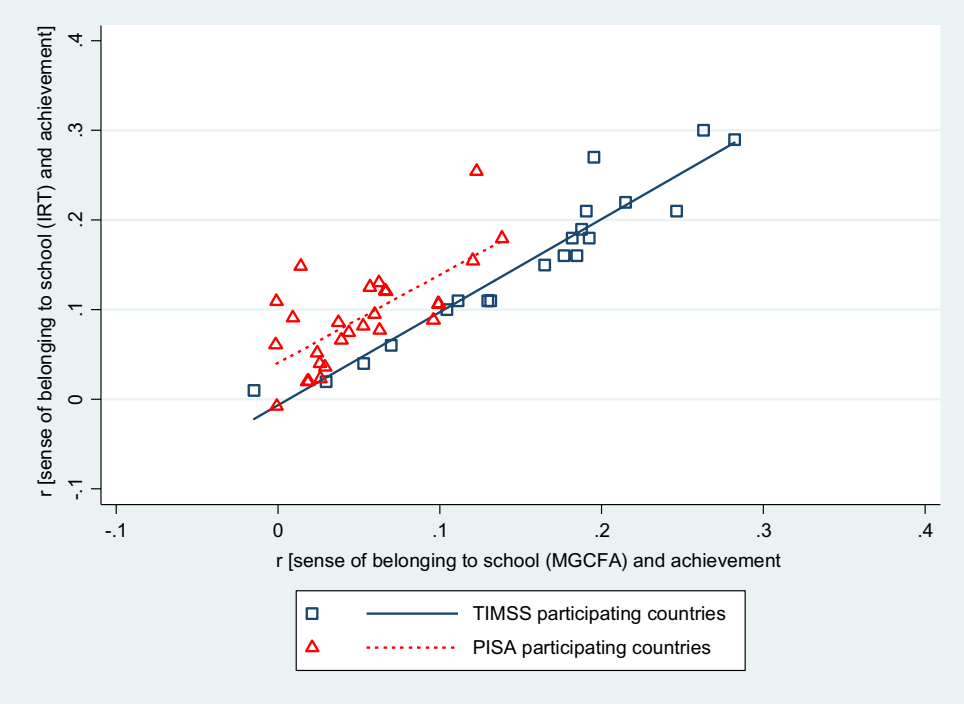


Figure 3. Relationship between within-culture correlations of science achievement and sense of belonging in school, where sense of belonging in school was scaled based on the MGCFA metric invariance model (*x*-axis) and IRT-based survey reported scale scores (*y*-axis), respectively.

and enjoyment of science, although measured with different numbers of items and some similarity in wording, showed metric invariance. Factor scores of these scales from MGCFA and IRT scaling produced rather similar patterning of correlations with science achievement, indicating robustness of scaling methodologies. Furthermore, the correlations between these scales and science achievement are highly related across the two surveys, whereas cultures with lower (e.g. Chile and Argentina) and higher (e.g. Korea, Australia and Ireland) correlations can be identified in both constructs regardless of the survey administered. In contrast, sense of belonging to school showed poor fit at the metric level in PISA, and its correlation with science achievement seemed to correlate negatively with the corresponding associations of the TIMSS scale. We discuss the need to carry out measurement invariance testing and the practical implications on PISA and TIMSS noncognitive assessments.

As mentioned before, the measurement invariance issue in the noncognitive assessments of the large-scale international surveys has hardly been a focus until recently. There is no over-stressing the importance of checking measurement invariance before making comparative inferences. A classic example of lack of invariance in self-report scales is the paradoxical correlations of students' self-report motivation and achievement in large-scale assessments (e.g. Van de gaer, Grisay, Schulz, & Gebhardt, 2012). In all participating countries, students' self-report learning motivation tended to show a positive correlation with achievement, whereas when scores were aggregated at the country level and the correlation is computed between countries' average levels of motivation and achievement, a negative correlation was

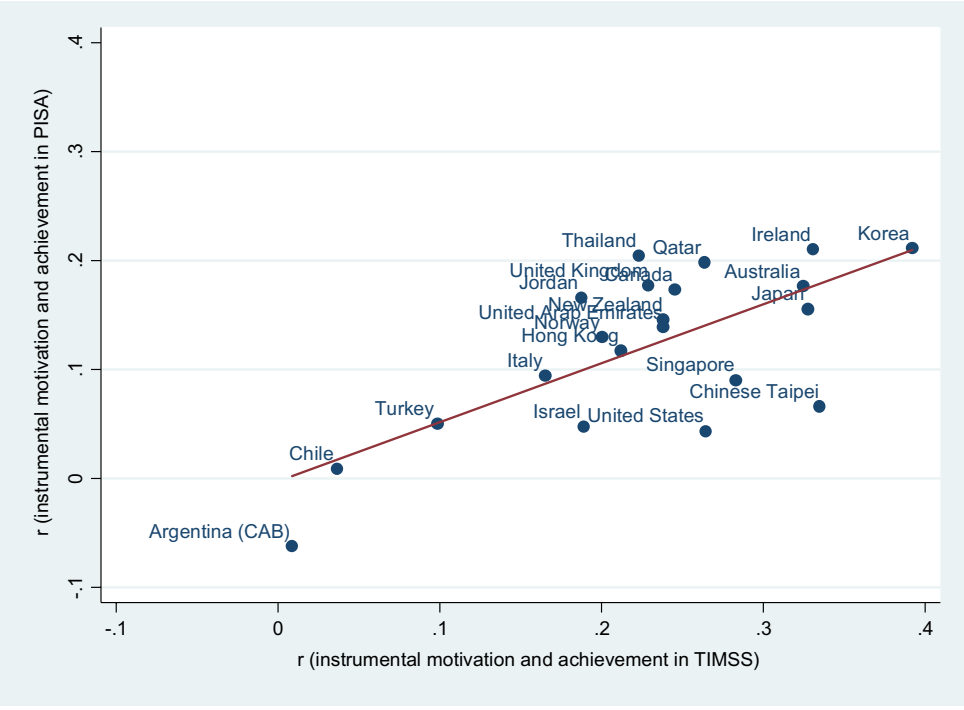


Figure 4. Relationship between within-culture correlations of instrumental motivation and science achievement in PISA and TIMSS.

Notes: Scale scores of instrumental motivation were derived from the metric MGCFA models, and scale scores for science achievement were those reported by the assessments.

found. That is, East Asian countries such as China, Korea and Japan, typically showing *high* scores on achievement, tend to have *low* scores on learning motivation. One possible reason for the paradox is measurement non-invariance caused by the different reference group effects (e.g. He, Buchholz, & Klieme, 2017): scale scores from different cultures should not be aggregated and compared, otherwise the patterning may be puzzling or erroneous. This may be the case for sense of belonging to school as measured in PISA 2015 in this paper. We encourage researchers utilising data of self-report Likert scales from different cultures to first assess measurement invariance, make it a routine as checking for reliability to form a sound psychometric basis, instead of assuming invariance and diving into comparative analysis. The present study suggests that MGCFA and IRT models produced rather similar factor scores; therefore, scale scores from both models can be used to analyse the relationship with achievement.

Zooming in the measurement invariance output of the current study, we find support of metric invariance of several scales, which forms a good basis for correlational comparisons. The fact that scalar invariance is not reached clearly warns researchers not to rank cultures based on these scales' mean scores. The hard-to-achieve scalar invariance with multiple cultures is understandable; however, we can either lower the strictness in parameter constraints to seek partial invariance (Byrne & van de Vijver, 2010) or approximate invariance (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2012) to tolerate trivial differences,

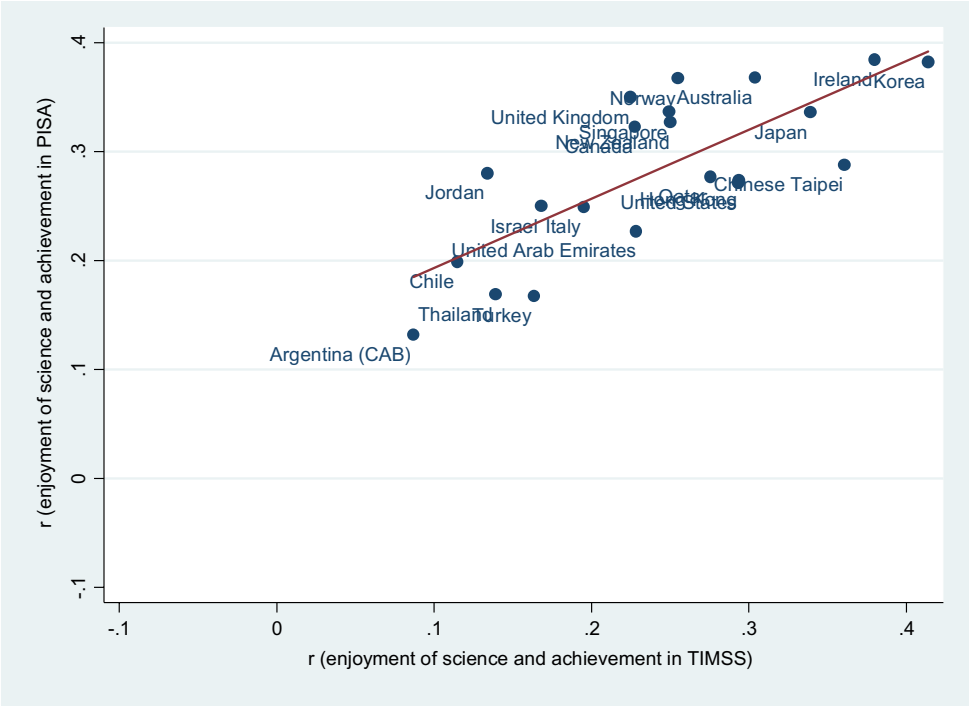


Figure 5. Relationship between within-culture correlations of enjoyment of science and science achievement in PISA and TIMSS.

Notes: Scale scores of enjoyment of science were derived from the metric MGCFA models, and scale scores for science achievement were those reported by the assessments.

or resort to response formats more resistant of bias (e.g. forced-choice format, situational judgement, ranking instead of rating, etc.) (Kyllonen & Bertling, 2014).

The cross-validation (particularly shown in Figures 4 and 5) provides confidence in ascertaining the positive associations between instrumental motivation and enjoyment of science with science achievement in most of the included cultures. These two constructs are more strongly related to achievement in cultures such as Korea, Ireland and Australia, and less strongly so in countries such as Argentina and Chile. It seems that the socio-economic development of the countries moderates the association between these constructs and achievement. This is confirmed by relating these culture-specific correlations with the Human Development Index (HDI) at the cultural level: the correlations are .26 and .46 for instrumental motivation in, and .73 and .63 for enjoyment of science in PISA and TIMSS, respectively. This moderation indicates that different intervention programmes may be appropriate in improving achievement. Whereas enhancing enjoyment and motivation may work better in the first cluster of countries, it may matter more to target other aspects for the second cluster of countries (such as resource allocation and teaching practice). When borrowing from successful examples from other countries, differences in cultural contexts should be taken into consideration.

To conclude, our study showcased the necessity in testing measurement invariance of scales before any cross-cultural comparisons of such scales, and how these scales prove to provide convergent (or lack of convergent) evidence when they are associated with science

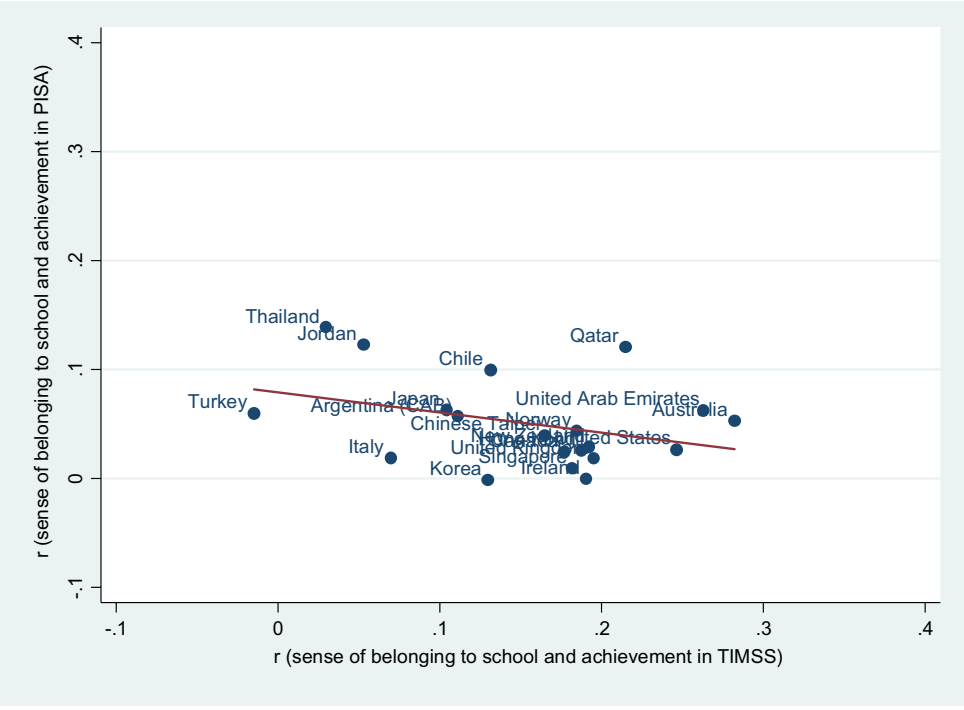


Figure 6. Relationship between within-culture correlations of sense of belonging to school and science achievement in PISA and TIMSS.

Notes: Scale scores of sense of belonging to school were derived from the metric MGCFA models, and scale scores for science achievement were those reported by the assessments. Lebanon was removed due to the negative value of the estimated variance for the latent factor of sense of belonging to school in PISA.

achievement in the two flagship large-scale educational surveys. Greater confidence is when the motivation and enjoyment of science constructs are used. We hope readers of the paper will include measurement invariance testing, either in the CFA or the IRT framework as a basic psychometric check and use data from both surveys to arrive at robust conclusions.

Notes

1. We use culture as a generic term; cultures in this paper indicates different groups being compared, or participating countries and economies in the two surveys.
2. The model did not converge with increased number of iterations, nor with the removal of cultures with negative covariance matrix.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Foundation Open Society Institute [grant numbers IN2018-41267 and IN2018-41261] and Marie Sklodowska-Curie Individual Fellowship European [grant Number 748788].

Notes on contributors

Jia He is post-doctoral researcher in the German Institute for International Educational Research, Germany. Her research involves data comparability in large-scale international surveys with innovative designs and sophisticated psychometric methods.

Fabián Barrera-Pedemonte is research fellow of the College for Interdisciplinary Educational Research (Germany) and the Center for Advanced Research in Education (Chile). His research interests include teachers' experiences, educational policies on teacher preparation, and cross-national quantitative comparisons for evidence-informed school system reform.

Janine Buchholz is post-doctoral researcher in the German Institute for International Educational Research, Germany. Her research interests include measurement invariance, modeling response processes in large scale assessments and multidimensional models of Item Response Theory.

ORCID

Jia He http://orcid.org/0000-0001-7310-4861

References

Ainley, M., & Ainley, J. (2011). Student engagement with science in early adolescence: The contribution of enjoyment to students' continuing interest in learning about science. *Contemporary Educational Psychology, 36*, 4–12. doi:10.1016/j.cedpsych.2010.08.001

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 495–508. doi:10.1080/10705511.2014.919210

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*, 497–529.

Boer, D., Hanke, K., & He, J. (in press). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*.

Braeken, J., & Blömeke, S. (2016). Comparing future teachers' beliefs across countries: Approximate measurement invariance with Bayesian elastic constraints for local item dependence and differential item functioning. *Assessment & Evaluation in Higher Education, 41*, 733–749. doi:10.1080/02602938.2016.1161005

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10*, 107–132. doi:10.1080/15305051003637306

Chiu, M. M., Chow, B. W.-Y., McBride, C., & Mol, S. T. (2016). Students' sense of belonging at school in 41 countries: Cross-cultural variability. *Journal of Cross-Cultural Psychology, 47*, 175–196. doi:10.1177/0022022115617031

Chiu, M. M., & Zeng, X. (2008). Family and motivation effects on mathematics achievement: Analyses of students in 41 countries. *Learning and Instruction, 18*, 321–336. doi:10.1016/j.learninstruc.2007.06.003

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: A cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology, 5*, 982. doi:10.3389/fpsyg.2014.00982

Cosgrove, J., & Cunningham, R. (2011). A multilevel model of science achievement of Irish students participating in PISA 2006. *The Irish Journal of Education / Iris Eireannach an Oideachais, 39*, 57–73.

Grabau, L. J., & Ma, X. (2017). Science engagement and science achievement in the context of science instruction: A multilevel analysis of U.S. students and schools. *International Journal of Science Education, 39*, 1045–1068. doi:10.1080/09500693.2017.1313468

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology, 48*, 319–334. doi:10.1177/0022022116687395

House, J. D. (2004). Cognitive-motivational characteristics and science achievements of adolescents students: Results from the TIMSS 1995 and TIMSS 1999 assessments. *International Journal of Instructional Media, 31*, 411–424.

Klieme, E. (2016). TIMSS 2015 and PISA 2015: How are they related on the country level? *DIPF Working Paper*. Retrieved from https://www.dipf.de/de/forschung/publikationen/pdf-publikationen/Klieme_TIMSS2015andPISA2015.pdf

Klieme, E., & Kuger, S. (2015). *PISA 2015 draft questionnaire framework*. Paris: OECD Publishing.

He, J., & Kubacka, K. (2015). *Data comparability in the Teaching and Learning International Survey (TALIS) 2008 and 2013*. OECD Education Working Paper NO 124. Paris: OECD Publishing.

Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (Eds.). (2016). *Assessing contexts of learning: An international perspective*. Cham: Springer.

Kyllonen, P. C., & Bertling, J. P. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–286). Boca Raton, FL: CRC Press.

Lam, T. Y. P., & Lau, K. C. (2014). Examining factors affecting science achievement of Hong Kong in PISA 2006 using hierarchical linear modeling. *International Journal of Science Education, 36*, 2463–2480. doi:10.1080/09500693.2013.879223

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016a). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 15.11–15.312). Boston College, TIMSS & PIRLS International Study Center.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016b). Creating and interpreting the TIMSS 2015 context questionnaire scales (Chapter 15). In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 1–312). Boston College, TIMSS & PIRLS International Study Center.

Michael, O. M., Ina, V. S. M., Alka, A., & Corinna, P. (2014). Context questionnaire scales in TIMSS and PIRLS 2011. In Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 299–316). Boca Raton, FL: CRC Press.

Mullis, I. V. S. & Martin, M. O. (Eds.). (2013). *TIMSS 2015 assessment frameworks*. Boston College, TIMSS & PIRLS International Study Center.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313–335. doi:10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Nagengast, B., & Marsh, H. W. (2014). Motivation and engagement in science around the globe: Testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006. In Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 317–344). Boca Raton, FL: CRC Press

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments*. U.S. Department of Education Economics: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch

OECD (2015). *PISA 2015 assessment and analytical framework*. Paris: OECD Publishing.

OECD (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris: OECD Publishing.

OECD (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*, 31–57. doi:10.1177/0013164413498257

Ryan, R. M., & Deci, E. L. (2009). Promoting self-determined school engagement: Motivation, learning, and well-being. In K. R. Wenzel & A. Wigfield (Eds.), *Educational psychology handbook series. Handbook of motivation at school* (pp. 171–195). New York, NY: Routledge/Taylor & Francis Group.

Stapleton, L. M. (2014). Incorporating sampling weights into single- and multilevel analyses. In Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 363–345). Boca Raton, FL: CRC Press.

Topçu, M. S., Erbilgin, E., & Arikan, S. (2016). Factors predicting Turkish and Korean students' science and mathematics achievement in TIMSS 2011. *Eurasia Journal of Mathematics, Science and Technology Education, 12*, 1711–1737. doi:10.12973/eurasia.2016.1530a

Van de gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology, 43*, 1205–1228. doi:10.1177/0022022111428083

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70. doi:10.1177/109442810031002

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*, 33–51. doi:10.1177/0022022100031001004

Wigfield, A., Eccles, J. S., & Rodriguez, D. (1998). The development of children's motivation in school contexts. *Review of Research in Education, 23*, 73–118.

Wu, M. (2010). *Comparing the similarities and differences of PISA 2003 and TIMSS* (OECD Education Working Papers, No. 32). Paris: OECD Publishing.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*, 1–26.

Yu, C. H. (2012). Examining the relationships among academic self-concept, instrumental motivation, and TIMSS 2007 science scores: A cross-cultural comparison of five East Asian countries/regions and the United States. *Educational Research and Evaluation, 18*, 713–731. doi:10.1080/13803611.2012.718511